LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 2, No. 2 (December 2008), pp. 212-243
http://nflrc.hawaii.edu/ldc/

Capturing Chaos: Rendering Handwritten Language Documents

John Henderson
University of Western Australia

This paper demonstrates how the nature of a source language document, and the broad goals set for the usability of the content, can direct the process of creating digital language documentation from that source. Gerhardt Laves's handwritten 1931 field notes on Noongar language and culture of southwestern Australia were retranscribed using an XML markup scheme and processed in various ways using XSLT. The central goals were to produce usable resources for community language activities and for linguistic and other scholarly analysis. A specific requirement for a rough facsimile representation, in recognizing that some of the graphic form of the notes was content that should be represented in the markup, contributed significantly to the specification of the markup scheme. Consultation with the Noongar community led to the recognition of Noongar families' rights in the materials and the recognition of culturally sensitive content, which together led to a requirement for multiple versions with varying content. The general nature of these handwritten notes also raises important issues of reliability and attribution that must be handled in the markup scheme.

1. INTRODUCTION[1]**.** The trend in digital language documentation is towards greater standardization, but within this trend, individual cases have their own implementation requirements. The methods must be motivated by the goals of each venture and must be tailored to the nature of the specific data. This paper demonstrates how the nature of a handwritten language document, and the broad goals set for the usability of the content, direct the process of creating digital language documentation. Further factors must be the technical expertise and resources available, but these points will not be addressed here. This paper identifies the key issues in the digitization of Gerhardt Laves's handwritten 1931 field notes on Noongar language and culture. These materials represent a valuable linguistic resource for an Australian Indigenous language that has been driven perilously close to extinction. Both the issues and the solutions in this case are relevant to a wide range of language documentation projects.

The goal of the Laves Digitization Project is a digital retranscription of the handwritten notes into a form that facilitates multiple uses of the content. The initial use is to produce a rough facsimile presentation in order to facilitate direct access to the content,

[1] The work reported here was largely conducted under contract to the Australian Institute for Aboriginal and Torres Strait Islander Studies (AIATSIS). I thank Peter Veth, Patrick McConvell, and David Nash of AIATSIS for their co-operation and assistance. I thank the members of the Noongar community reference group established for this project, my co-researchers (Hannah McGlade, Kim Scott, and Denise Smith-Ali) and the research assistants who did much of the re-transcription (Andrew Gargett, Denham Harry, and Harry Wykman, who also assisted in developing the XSLT used in this project). Other aspects of this project have been reported in Henderson et al. 2003, Henderson 2006, and Scott et al. 2006. Thanks to Ezzard Flowers for permission to reproduce parts of a text by Freddie Winmer.

E-ISSN 1934-5275

which is often obscured by the nature and physical condition of the original field notes. Other important uses are linguistic analysis and the republication of the traditional texts that Laves recorded. Three aspects of the notes have particularly influenced how the retranscription is implemented. First, there are different rights in the materials and culturally sensitive content that necessitate multiple versions of the notes. Second, the nature of the notes raises many issues of reliability in the retranscription. Third, allowing for (even a rough) facsimile representation requires recognizing the graphic form of text on the page as part of the content to be captured. The issues of reliability and the graphic form of a text are of course well-recognized in the tradition of European textual scholarship, especially as relates to historical manuscripts (Greetham 1994). This tradition offers a useful background to the present project, not least because many of the issues that have arisen in that tradition are addressed in the *Guidelines for electronic text encoding and interchange* (Burnard and Bauman 2007, also known as the TEI P5 Guidelines).

The discussion is organized as follows. In section 2 I describe the nature of the field notes and the language they record. The goals and general parameters of the digitization project are described in section 3, and the details of their implementation in section 4.

2. LAVES'S NOONGAR FIELD NOTES. Gerhardt Laves (1906–1993) was an American graduate student who conducted field research on a number of Australian languages from 1929 to 1931, but did not complete his research to publication after he returned to America. He published only two brief notes on his work, but produced a vast collection of field notes (and a few wax cylinder audio recordings of a language from the northwest of Australia). Laves kept his Australian field notes for over fifty years, largely unknown to Australian researchers and the language communities involved. They only became public in the mid-1980s, when copies were obtained by what is now the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS). The more readable originals were deposited in the AIATSIS archive after Laves's death. See Nash 1993 for further biographical information.

Laves's work is particularly valuable today for two reasons. First, the quality of his description is much better than that of many other early researchers. He was the first researcher of Australian languages who had professional training in linguistics, having been taught by Edward Sapir. He was also hosted in, and prepared for, his Australian work by the noted anthropologist A.R. Radcliffe-Brown. Second, his more detailed studies were done at a time when more people spoke these languages fluently. Today many of the languages that Laves studied have moved closer to extinction.

Laves conducted major studies of six languages, and minor studies of more than twelve others. One of his major studies was conducted at Albany in the south of Western Australia on varieties that he identified as Kurrinj2 and Minong, today described as dialects of

[2] The notes contain three variants of this name, retranscribed as: *Kurrinj* (where "rr" represents a tap/trill), *Kurinj* (where "r" represents an approximant), and *Kurin*, which occurs only as a stamped label in the top right corner of the loose slips. See figure 2. While the last spelling does show up in at least one other source, it is likely that it results here because the stamp Laves used appears to have permitted a maximum of only five characters. This is consistent with the corresponding stamped labels in his notes on other Australian languages.

Noongar (also Nyungar, Nyoongar, and other variants, partly reflecting different spelling practices and partly different pronunciations). The Noongar materials are of two types—notebooks and loose slips. There are ten notebooks, 1570 pages in total. These mostly contain stories told to Laves in Noongar by a range of speakers, together with partial interlinear glossing and a loose translation that Laves called a résumé. There are also some other notes including genealogies of the speakers' families (which he presumably collected in order to investigate the kinship system). An example is given in figure 1 below. Apart from the notebooks, there are another 2,453 loose slips with vocabulary and extra notes on some of the language in the notebooks. An example is given in figure 2 below.

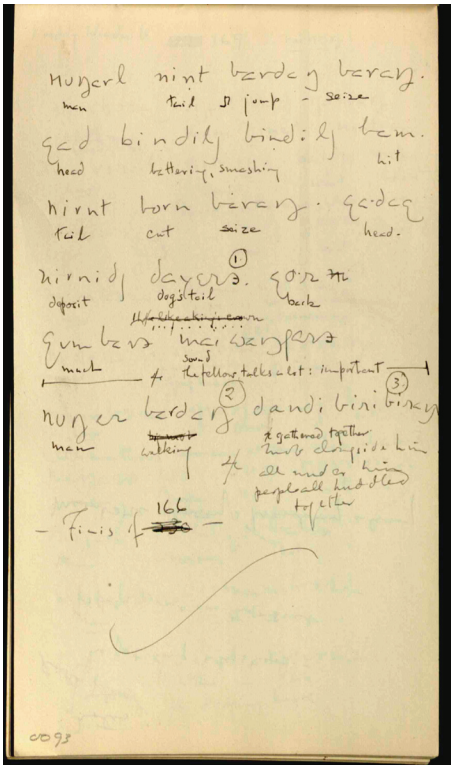

FIGURE 1: Sample page from the notebook volumes.[3]

[3] This page is reproduced here by kind permission of Ezzard Flowers on behalf of the family of Freddy Winmer, the author of the text from which it is taken.

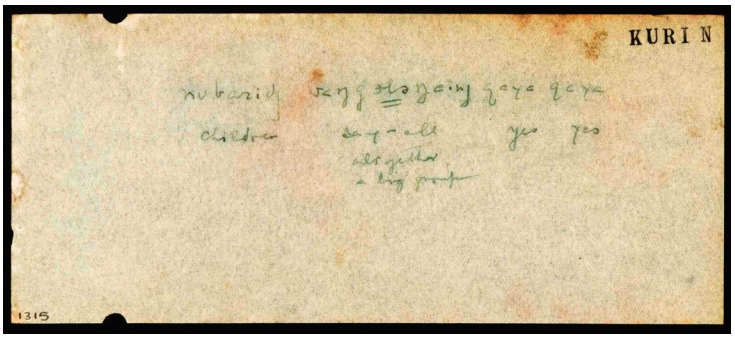

FIGURE 2: Sample loose slip.

The Noongar language is now in very limited everyday use and is severely endangered. There are many Noongar people today who have command of at least some expressions of the language, but there are likely to be relatively few more knowledgeable speakers. There do not appear to be any speakers who communicate over the full range of everyday functions of language using Noongar.

In the last couple of decades there have been numerous small-scale contributions to revitalization of the Noongar language. Programs, based mostly on second-language teaching methodology, are offered in a number of schools in the region. There is a small range of supporting materials, including published texts and language-learning resources. No extensive grammatical descriptions or detailed lexicographical works have been published. The works that have been published focus on lexical information and give very basic treatments of the sound system and grammar. None appear to draw on Laves's work. There are two substantial compilations of lexical data from historical sources, Bindon and Chadwick 1992 and Alan Dench's 1980s lexical database. Both mostly preserve the original written forms and do not provide reconstitutions of actual or likely phonological forms and meanings on the basis of the historical descriptions. Dench 2000 discusses methods of producing such reconstitutions for Noongar. Dench 1994 gives a multi-dialectal wordlist, which consists largely of such reconstitutions from historical sources. Douglas (1968) and von Brandenstein (1988) added their own field data to lexical data from earlier historical sources, with the former giving a basic sketch of the grammar. An unpublished grammatical description was reportedly produced by Francesca Merlan in the 1980s.

The traditional Noongar area is large, around 193,000 square kilometers. There was almost certainly dialect variation before European settlement began in 1826, but no in-depth studies of the nature of the variation have been published. O'Grady et al. 1966 lists thirteen dialects. Dench (1994), following unpublished work by Morphy, recognized three major dialects—northern, southern, and southwestern—differing "mainly in their varying pronunciations of similar words" (1994:174). He also gave a more fine-grained division into six dialect regions. Laves worked with speakers from the south, and predominantly the mid to eastern part of the south. His work is the major source on the dialects of that area, within which he reported a number of dialect distinctions, though this is based on limited data from only a handful of speakers.

The present state of the Laves materials is a cogent lesson on the importance of proper practice in language documentation and professional archiving. Laves stored the notes in his various homes over the years, at least sometimes in a basement or attic. Many of the loose slips have been damaged by water, mildew, and the ravages of time. Some slips are now completely unreadable, and many require some educated guesswork where the handwriting is obscured or faded. Fortunately the notebook volumes with the original texts are in better physical condition.

The Laves materials are all handwritten (with the exception of stamps at the top of the loose slips to indicate language/dialect), and, as is often the case, there are frequent problems interpreting the handwriting, especially when a page is also in poor physical condition. Naturally, this is partly a matter of developing some familiarity with Laves's handwriting. A researcher who has worked extensively with the originals and developed some familiarity has an advantage over the casual browser, but still faces difficulties.

Describing a work as field notes implies a range of possibilities from the extensive, detailed, and transparent to the cryptic and minimal, of which the latter might normally be expected to be of use only to their author, as aids to the memory of the language event. Fortunately, Laves's notes generally tend towards the explicit (with the most significant exception being his use of idiosyncratic symbols). But of course, it is not reasonable to expect every item in field notes to reflect their ultimate analysis. For example, the gloss given for a specific instance of a word in field notes may be only a small part of the evidence of the meaning of a lexical item as it would be expressed in a dictionary.

As with other rough or raw notes, interpretation often relies not only on the written words themselves but on their position on the page. Interlinear glossing is a common case where relative position is significant, but it is a well-established standard that can be readily interpreted in terms of content type: glossed element and its gloss. However, the notes abound with over-written corrections, deletions, insertions, sparse marginal comments, and (fragments of) notes whose interpretation depends on their placement on the page relative to the element they are annotations to. The notes also make frequent use of graphic devices, such as lines, arrows, and enclosing boundaries, to draw attention to some element or to indicate a relation between two elements. Because they are raw notes, it is not surprising that the positioning or graphic devices are not always used consistently.

This project is not merely a technical exercise in research archiving on the part of AIATSIS. It has involved engagement with the Noongar community and, because Laves had identified the person who provided each text in 1931, it has involved engagement with the individual families of his original sources. The materials have great cultural importance for these families, especially because the passing down of language and cultural knowledge in the region has been seriously affected since the arrival of Europeans. The various families claim rights in the materials, especially a right to decide how they are managed now in terms of access and use. Consultation with the community resulted in an extensive protocol document (Scott et al. 2006). Key issues from community consultation for the technical aspects of the project were usability of the digital version for community members, restrictions on access to texts to the specific families of the original source, and restrictions on other culturally sensitive content.

3. GOALS AND PROJECT PARAMETERS. AIATSIS initiated the project to digitize the Noongar materials in 2002, first making digital photographic images of each page and then having the content retranscribed so that it could be made available in more accessible forms.4 Following the developing standards and principles being applied to language documentation and description (Bird and Simons 2003), it was decided that the digital retranscription would use a descriptive markup scheme implemented in XML.

The primary goal was that the retranscribed version of the field notes would be flexible enough to be readily put to multiple uses for various users over time. Two immediate types of users were envisaged: (1) members of the Noongar families of Laves's sources in 1931, together with members of the wider Noongar community, and (2) academic researchers in linguistics and in related disciplines such as anthropology. These two audiences (which are not strictly mutually exclusive) have overlapping interests but tend towards different focuses.

There are certain consequences for the XML retranscription that follow from this goal of allowing for multiple uses. First, there is an advantage in preserving as closely as possible the original content, rather than substituting editorial interpretation for original content. Editorial interpretation can be added at any time as an additional layer of information in the retranscription where necessary. Some aspects of the notes are difficult to interpret definitively, and if the primary goal is to allow for multiple future uses, then future users will need to be able to make their own interpretations. Speculative interpretations are minimized in the retranscription (or at least indicated as such in a transcriber's note) and any ambiguity, vagueness, or other such complexities are preserved for later interpretation in the course of some specific use. Corrections, insertions, etc. are preserved, and alternative retranscriptions are recognized where Laves's intention is not clear. There are many instances where Laves's handwriting is not clearly legible and where the reliability of the retranscription therefore cannot be guaranteed.

Second, there is an advantage in representing some of the graphic form of the original pages in the XML retranscription, because the retranscription can then be used to produce facsimile representations of the pages.[5] This has advantages in relation to perceptions of the validity and reliability of the content since these will depend to some degree on a transparent relationship between the facsimile version and the originals. There are three reasons for this. First, there was a perception that some community members would prefer the immediacy of experiencing the notes in facsimile form in order to have greater

---

[4] The digitial images were supplied from AIATSIS in TIFF format (600DPI; RGB; typical image size: 2749 x 4540; file size around 28Mb). For the distribution packages described below, the images were converted to a compressed form in JPEG format (600DPI; RGB; image size 1475 x 2453), which at approximately 120 kB per image allowed the entire set of images, and accompanying XML and HTML documents to be copied to a single CD-ROM.

[5] The term *facsimile (representation)* is used here to describe an attempt to represent as closely as practical the visual properties of the page, especially the layout of the text components. In the categories recognized in the textual scholarship tradition, this form of representation is perhaps closer in intent to the *print facsimile* than to the *diplomatic transcript*. See Haugen 2008 and Greetham 1994.

confidence that they are valid representations of the originals (while maintaining a higher level of readability than the originals). Any lack of confidence in a heavily edited version might send those people back to working with the original version, thereby defeating part of the purpose of the exercise. Second, some aspects of the form are in this case part of the content. There are many aspects of the micro-structure of pages whose interpretations rely in part on their position on the page, for example: glosses, insertions, corrections, author's notes, etc. To the degree that the initial facsimile version can match the relative positioning of items on the page, the same interpretations are available to the reader. The positioning could certainly be described in words in the retranscription, but this alone will not always facilitate the same interpretations as a graphic representation. Third, given all the complexities in interpreting the notes, there is a clear advantage in having easy recourse to (images of) the originals in order to compare them to the retranscription. Even if the form of a facsimile version only roughly matches that of the original (or a photo reproduction), it still facilitates comparison by allowing the user to more easily locate the positions of relevant items in the two representations.

The result is that the content and form of the original notes are best represented in the initial readable version as a facsimile representation with at least a rough correspondence to the original page. This is used most effectively when presented in a parallel view with an image of the corresponding page, as shown in figure 3. Allowing for this type of use dictates some of the specifications for the XML tagset for this project.

The initial facsimile version was of the *entire* set of notes and was produced to facilitate further uses of the XML retranscription. Most important, this was necessary as an initial step so that senior Noongar people could conduct a detailed assessment of the content in order to identify any culturally sensitive content that might require access restrictions. Various different restrictions were placed on parts of the texts by family representatives.

For a digital resource that is intended for a wide range of users, it is necessary to take a broad view of usability. First, what application will be available to present the resource for reading? In this case, an obvious way to implement a readable version for a variety of users is via browser software. Second, beyond just reading, users commonly expect the basic facilities found in modern document-handling software. Users will probably expect at least to be able to browse the document and search in it. These requirements are not difficult to meet in browser software, at least in simple ways, but even something as basic as copying text from the browser window into a word processor document for further use can present problems (in addition to any loss of style and format information).[6]

[6] For example, if a section of text contains certain non-Latin characters, including [ɣ], it is not possible to copy the section from the browser version viewed in *Firefox* (v.2.0.0.4) and to paste any of that section into a *Microsoft Word* document (*Word 2004 for Mac*, v.11.3.5). This is despite the fact that *Word* is Unicode-capable and the offending characters can otherwise be successfully entered via the keyboard, and copied and pasted between *Word* documents. The identical section copied from *Firefox* can, however, be successfully pasted into a *TextEdit* document.

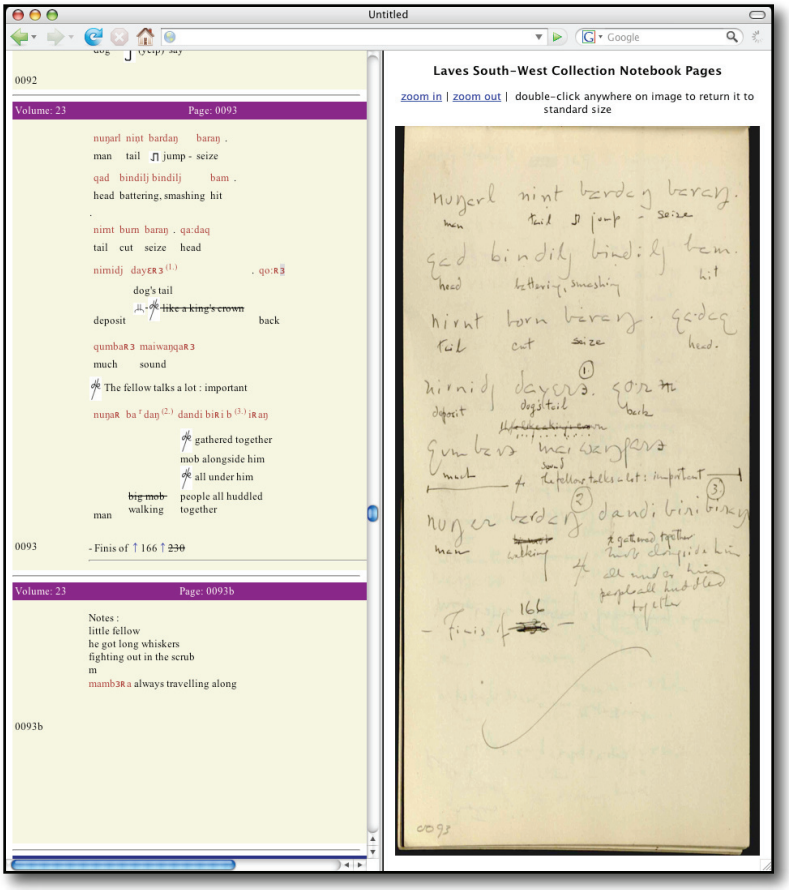

FIGURE 3: Parallel presentation of page image and facsimile retranscription.[7]

For some users, though, any kind of digital version will be of limited value. With the original paper document, the cost of making copies of the more than 4,000 pages, not to mention the unwieldiness of that much hardcopy, had clearly affected use of these materials in the past. Very few individuals or institutions held copies; fewer made use of them. A digital version has the benefit of being easy and cheap to distribute, as well as the benefits discussed above, but it has limitations.[8] Not all members of the Noongar community, es-

[7] These pages are reproduced here by kind permission of Ezzard Flowers on behalf of the family of Freddy Winmer, the author of the text from which they are taken.

[8] It also complicates matters in one sense because the greater accessibility of the digital versions accentuates the issues of restricted access to culturally sensitive content. The inaccessibility of the original form of the notes tended to limit all access.

pecially older people, have access to computers or familiarity with using them, and will generally prefer some form of hardcopy. Even people who are very familiar with digital technology often prefer hardcopy to the screen for reading large amounts of text. This requirement can, of course, be easily met by printing out a hardcopy from the browser, provided that this can be done in an appropriate and compact format.

Community use is also facilitated by minimizing the level of technical skills required to access the digital versions. For this reason, the browser-viewable versions that are distributed on CD-ROM are technically simple, relying on HTML, limited use of Javascript, and the image files themselves. They do not require any special plug-ins to view images or text, unlike for example the *Luce Papers* project[9] which would appear to be primarily directed toward a scholarly audience.

The initial facsimile version of the entire set of notes is not, of course, the only use to be taken into account in designing the XML implementation. It must also be possible to produce other facsimile versions, and to allow automatic processing of different kinds, including information retrieval. Examples include sophisticated searching for the purposes of grammatical analysis, and the automatic extraction of word and corresponding gloss information for the purposes of developing a glossary or more extensive lexicographic work.

The initial retranscription task is somewhat simplified because Laves provided relatively little grammatical analysis, which is perhaps not surprising in field notes. His interlinear glossing is mostly not in morpheme-by-morpheme form; there is little morphological analysis. More extensive grammatical annotation (part of speech, morphological categories, phrase structure, etc.) can, and, one hopes will, be added at some stage, but this will first require more grammatical analysis of the language.

In the following sections, I discuss the issues presented by specific features of the originals.

4. IMPLEMENTATION. The central point of this paper is how the implementation follows from both the intended goals and the nature of the original data. The focus of the implementation here is on the design of the XML scheme that is used for the retranscription and the automatic processing of the XML for specific purposes. In this project, the processing is achieved mainly by using the Extensible Stylesheet Language (XSL) family. 10 The XML retranscription is processed using XSLT stylesheets and output either as XHTML to be viewed in a browser window or as another XML document with modified content.

For readers who are unfamiliar with these technologies, the following discussion should give some idea of their application in this context.11 XML is a general set of speci-

[9] http://www.sealang.net/archives/luce/

[10] http://www.w3.org/Style/XSL/. The term stylesheet is partly misleading because XSL is not limited to formatting XML documents for presentation.

[11] The enormous literature on these technologies includes many basic introductions and tutorials. A source relevant for using XML in linguistic work is Chapter 5, *A gentle introduction to XML* , in Burnard and Bauman 2007.

fications for creating mark-up languages for particular purposes. The key notion is that features of a text or other data can be described in annotations that are added to the text in the form of tags. For example, to indicate that a word in the retranscription is not clearly legible, it is preceded by an opening tag <unclear> and followed by the corresponding closing tag </unclear>, thus "<unclear>qarl</unclear>." Structural elements within a document, such as texts, pages, paragraphs, etc., can also be identified using tags, for example "<text>…</text>." A tag can also add further annotation by using attributes. For example, in this project the name of the author of a text is expressed within the opening tag as an attribute of the <text> element: "<text author="xyz">…</text>." XML does not itself specify the actual tags to be used (<unclear>, <text>, etc.). These are determined by users, who may design their own custom tagset, or use an existing general standard such as the TEI Guidelines or a special-purpose scheme based on TEI such as the EpiDoc scheme for epigraphic documents.12

Using XSLT stylesheets to process the content of XML documents can be illustrated by the template in (4). This converts any parts of the retranscription that occur within <unclear>…</unclear> tags in the XML document into an XHTML with the style attribute specified so that those parts of the text appear not between the <unclear> tags but with a light grey background when viewed in a browser window. Other features of XSLT facilitate more sophisticated processing. The output of a given XML element can be made conditional on the text content within that element or on the values of any attributes the element may have. For example, to output the texts of a particular author in a distinctive way, the XSL template specifies both the XML element and its author attribute, as illustrated in (5). The output specified by that template then applies only to texts by that author. If an XSL template does not specify an output for a given XML element, then the content of that XML element is effectively suppressed in the output. This, for example, allows outputting *only* the texts of a particular author (which is necessary to produce some of the versions discussed below). Different XSL stylesheets can produce different kinds of output from the same XML document.

(4)

```
<xsl:template match= "unclear">
          <span style="background-color: lightgrey">
                <xsl:apply-templates/>
          </span>
</xsl:template>
```

(5)

```
    <xsl:template match="text[@author="xyz"]"> … </xsl:template>
```

Determining the optimal XML markup scheme for a project such as this is a major task. A key initial decision was not to follow the TEI Guidelines, at least directly. Instead a

[12] http://epidoc.sourceforge.net/

markup scheme was devised from the ground up to closely fit the nature of the project, but informed nonetheless by the general principles and classifications in the TEI scheme. This decision was influenced by a number of factors. First, the TEI Guidelines were perceived to be very complex.13 An obvious problem is that such an extensive scheme has a large number of tags to learn to use, but a more subtle problem was that the definitions and exemplification of the tags in the Guidelines were in some cases not enough to give the transcribers confidence that their use of the tags would be comparable with other projects. Detailed technical descriptions of comparable projects, with extensive exemplification, would no doubt have reduced the effect of this factor. Second, at the time there were no tools apparent that made use of the TEI scheme in a way that offered any immediate advantage for this project, for example editors and generic XSL stylesheets.

In retrospect, there are both positive and negative consequences of the decision to use a largely custom tagset. A clear disadvantage of not using TEI is that the documents are less comparable with other works and therefore less able to benefit from the advantages that a standard offers. This means that there will be additional work for the creators of any future tools to manipulate the documents, such as XSL stylesheets, no matter how well documented the custom tagset is. On the positive side, designing the tagset from the ground up had the advantage that it closely matches the specific requirements of the source documents and was therefore efficient to implement when keyboarding the transcription. Balanced against the advantages of using a standard scheme, there is an apparently perverse, and possibly controversial, advantage in using a custom tagset which is *less* amenable to any future generic manipulation, say by generic TEI stylesheets. The advantage is that this reduces the chances that culturally-sensitive content will become separated from the annotations which specify restrictions on access. (See discussion below.)

Overriding this comparison of the relative advantages of the two approaches, however, is the fact that the custom scheme is largely conformable with the Guidelines in that many of the tags used are directly replaceable by TEI tags, and others could be converted to TEI by some relatively simple processing (at least in theory), though in other cases some manual editing would certainly be required. The discussion below is not intended to provide a detailed comparison between the custom tagset used and the nearest TEI equivalents, but it does include some occasional comments.

In the remainder of this section, I discuss the custom XML markup scheme in five areas: core structure, document structure, multiple versions, reliability, and graphic form as content. The discussion identifies these key features of the scheme but does not exhaustively describe all of the tags used. The XML retranscription of the page in (1) and (3) is given as an illustration of the markup scheme in the Appendix.

4.1 CORE STRUCTURE. The basic aspects of the markup scheme are those that represent the actual Noongar language data, especially interlinear glossed text and vocabulary definitions. Individual words or sections in Noongar that are not part of structured glossing or definitions are marked as <nyungar>, (which can be thought of as an abbreviated equivalent of something like <section language="nyungar">). Structured definitions, which occur

[13] The full version of the current TEI P5 Guidelines (Burnard and Bauman 2007) is 1295 pages. Even the Lite version (Burnard and Sperberg-McQueen 2006) prints out at 166 pages.

mostly in the slips, are represented as a set that consists of the element being defined and the definition given for it, as illustrated in (6).

(6)

```
<def_set>
       <defd>dandan</defd>
       <defn>one behind another</defn>
</def_set>
```

The structure for interlinear glossing given in (7) is centered on the element being glossed and the gloss Laves gives for it, marked as <glossed> and <gloss>. In the quite common case in the Laves texts where no gloss is given for a particular item, it is represented with empty tags: <gloss></gloss>. A line of glossed elements and the corresponding line of glosses form a glossing line pair, <gl_pair>. This strategy differs from other sources such as Bow et al. 2003 where the individual glossed and gloss items form a unit, and a line consists of a number of such units.14 The glossing line pair approach was originally chosen in the present project because it corresponds more closely to XHTML table structure and could permit simpler transformation for presentation in XHTML. The relationship between the glossed and gloss item is not lost, however, because the two strategies are readily interchangeable, provided that there is a <gloss> element corresponding to each <glossed> element. In fact, matching of individual <glossed> and <gloss> elements has been done in this project, using XSLT to extract such pairs in order to produce wordlist data. A gloss set (<gl_set>) contains one or more glossing line pairs (<gl_pair>) and one free translation (<g_translation>) if one is given. A text proper consists of one or more gloss sets.

(7) Basic structure for inter-linear glossing.

```
<gl_set>
       <gl_pair>
              <glossed_line>
                     <glossed>…</glossed>
              </glossed_line
              <gloss_line>
                           <gloss>…</gloss>
              </gloss_line>
       </gl_pair>
       <g_translation>…<g_translation>
</gl_set>
```

4.2. DOCUMENT STRUCTURE. Since the retranscription scheme is intended to allow for facsimile representation, the page image must be a key element of the master XML documents. Each document includes a set of <image> elements, and the entire document is ac-

[14] Bow et al. (2003) see a specific value in this approach because automatic line-wrapping can be readily implemented in a way that preserves the glossed-gloss formatting—that is, line-wrapping does not apply separately to the glossed line and the gloss line.

cordingly labeled an <imageset>. Each <image> element represents a whole page (except in certain cases discussed below). The master XML document that contains the transcriptions of the loose slips is structured slightly differently from the master XML document that contains the volume pages. For the loose slips, the page images are not organized into any groups. Notebook pages, however, are grouped into the ten notebook volumes, which are represented as <vol> with the number attribute representing Laves's own numbering of the notebooks from 18 to 27.

Most of the content of the notebook volumes is traditional texts in Noongar and/or English. Each is represented using <text> with the id attribute representing the labeling scheme used by Laves—for example "121QQ," and the author attribute representing standardized versions15 of the names of the Noongar consultants who provided the text (or in a few cases, "author not indicated"). Each text may consist of a Noongar language part, usually with partial interlinear glossing, and/or a loose English translation or summary that Laves refers to as a "résumé," represented as <text_proper> and <resume>. The <text> element can also contain other material, mostly Laves's notes on the text. The general <text> structure is illustrated in (8). There is no specific ordering between the <text_proper>, <resume>, and any other element within a <text>. Moreover, the text may be discontinuous in that the text proper is in one location in the notebooks and the résumé in another, or, for example, because some nontext material is interposed within a text proper or résumé. In that case, the discontinuous parts are represented in separate <text> structures that are unified by sharing the same id value. Images that are not part of Laves's text units are represented as <nonText>.

(8) Basic document structure of the notebook volumes master, showing text units.

```
<imageset>
       <vol number= "X">
              <nonText>
                            <image>…</image>
                            <image>…</image>
                            …
              </nonText>
              <text id="…" author="…">
                     <text_proper>
                            <image>…</image>
                            <image>…</image>
                            …
                     </text_proper>
```

[15] Standardized forms of the names are useful here because in the actual texts Laves sometimes gives different versions of some authors' names. These involve minor spelling differences, apparently different anglicizations of Noongar names, and in one case, a playful Latinization. For example, "Freddy Winmer" = "Freddy Windmill" = "Fridiricus Windmillii." The different variants of the authors' names are documented outside the main XML documents. The representation of these variants within the transcriptions would be improved by associating the standardized form of the name with each instance of the name in whatever variant form.

```
                        <resume>
                              <image>…</image>
                              <image>…</image>
                              …
                        </resume>
                  </text>
                  …
            </vol>
            <vol number= "X+1">
                  …
            </vol>
            …
      </imageset>
```

The representation of text units also presents a more fundamental problem that arises from the constraints of XML. A single text part (text proper or résumé) may start in the middle of a page below some other material and/or end in the middle of a page above some other material. This means that although a text can extend over a number of pages, a text element cannot simply be a grouping of whole pages. The problem is that text elements and pages can overlap, and overlapping structures are not permitted in XML, which is strictly hierarchical. A number of methods have been developed to overcome this limitation of XML (Burnard and Bauman 2007). The fragmentation method is used in this project. If the text starts or ends in the middle of a page with other content, the page is transcribed as two separate <image> elements that are unified by sharing the same page number, etc. The status of these part pages is indicated in the <image> element by the *incomplete* and *continued* attributes.

The edited example in (9) shows the markup scheme for a text in which the text proper starts at the beginning of page 58 and extends until halfway down page 65. The résumé then follows immediately on page 65 and finishes at the bottom of that page. In order to represent the overlap of the text proper, page 65 is split over two <image> elements. The first <image> element includes only the final part of the text proper and is immediately followed by the closing </text_proper> tag. The *incomplete* attribute indicates that the retranscription of the page is not completed within that <image> element. The second <image> element contains only the résumé that completes page 65, and it is enclosed within <resume> tags. The *continued* attribute of the second <image> element indicates that this <image> element continues the transcription of a page.[16] This meets requirements because the hierarchical structure of the XML is preserved while achieving the desired functionality. The <text_proper> and <resume> elements can be independently extracted and processed for presentation, while the content of the two <image> elements can readily be merged for presentation of the page as a whole because they share the same page number.

[16] This scheme can also represent cases where a single page must be represented by *more than two* <image> elements. The intervening <image> elements are specified as both *incomplete* and *continued*.

(9)

```
<text>
      <text_proper>
            <image volume="18" page="58">…</image>
            …
            <image volume="18" page="65" incomplete="yes">
                  …
            </image>
      </text_proper>
      <resume>
            <image volume="18" page="65" continued="yes">
                  …
            </image>
      </resume>
</text>
```

4.3 VERSIONS. The value of XML and related technology for language documentation lies in part in its ability to automatically generate multiple presentations of the data or selections from it in different versions. There is a cost, of course, in the complexity of the processing task and long-term maintainability. The master version of the Laves retranscription is in just two XML documents, for the loose slips and notebook pages respectively, but the project produces a relatively large number of derivative versions.

The texts in the notebooks are, for the most part, attributed to specific Noongar authors from 1931.[17] Most texts have a single author; a few have multiple authors. Some fragmentary texts have no author attributed. The Protocol document recognizes that for each author there are various families today who have rights in the different texts, and that the rights are not held by the Noongar community as a whole over the set of texts as a whole. For this reason there is no overall facsimile version for Noongar community use, but different facsimile versions with the appropriate selection for the different families. In order to facilitate the production of these different facsimile versions, a derivative XML version is first produced from the master document with a separate XML document per author, regrouping the individual texts from throughout the notebook volumes. This derivative XML document is then separately processed using XSLT to produce a facsimile version of the texts of just that author.

The preparation of the appropriate materials for each family is further complicated because some texts have multiple authors, in which case more than one family may have rights in the text. And all the families have rights in a few fragmentary stories and other content where Laves does not identify the author. Thus there can be a many-to-many relationship between the 1931 authors and families today. A separate digital "package" of documents is prepared for each family. The distribution of documents to the overlapping packages is largely automated to facilitate updating of individual documents.

---

[17] Authorship would appear to be the appropriate role for the Noongar language texts that purport to be a verbatim record.

The Protocol also deals with another key restriction on access, often described as culturally sensitive content. In general terms customary law may restrict some content to men only, and in some cases only to men with the appropriate affiliations. Some content may be judged by community members to be suitable for adults only. As part of the process of community consultation, the texts are vetted as much as possible by men with the proper authority from the different families. The restrictions they established were implemented in XML in two ways: at the page image level with the attribute specification *has_restrictions="yes"*, and at the specific element level with tags that identify the domain and authority of the restriction with the attributes shown in (10). The attribute *restriction_type* currently can only have the value "closed."

(10)

```
        <restriction
        restriction_type="closed"
        restriction_authority=   "(person   authorising   the
  restriction)"
        restriction_date="…" />
…
</restriction>
```

For facsimile versions for community use, two separate versions are produced of the texts by a given author—a restricted version and a version for general use by family members. In both versions, the header for a restricted page includes the label RESTRICTED CONTENT, but in the family version, the actual content for that page is suppressed, as illustrated below.

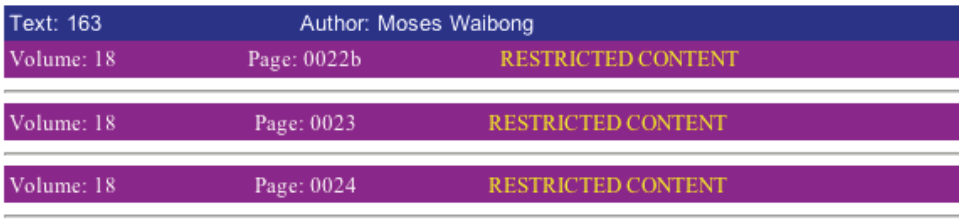

FIGURE 11: Suppression of restricted content in facsimile presentation. Compare with FIGURE 3.

To date, the texts of only one author have been formally vetted in this way. All other material is labeled as *NOTCLEARED* to indicate that it may or may not contain restricted material. Given the cultural importance of this three-way classification, it has to go beyond these facsimile versions. In order to minimize the chance that the master XML documents or any XML or XHTML version might accidentally be made available contra the restrictions, every single document has the appropriate restriction label incorporated in the filename, for example *VolumesMaster.notcleared.xml*, *Moses Waibong.restricted.html*, and *Moses Waibong.family.html*. Access restrictions of these types could also be implemented at some point in the future by using a permissions system as found in database management

software. Such software has not been used in the current digital versions for community use because these versions are intended to minimize the technical requirements on users. At this stage, no content has been cleared for open use—that is, outside the respective Noongar families—though this might be expected to happen at some point in the future.

The combination of the restriction types and the different selections from the master documents, with versions in both XML and XHTML, results in a large number of version documents. This has important consequences for maintaining the integrity of the content over all versions. Given that issues of legibility and interpretation of the handwritten originals can be expected to result in a reasonably high error rate in retranscription, any errors that are detected have to be corrected in all version documents. The same is true for any alternative transcriptions or annotations that a user may want to add. The necessary updating is for the most part achieved automatically by a series of scripts that generate the various documents from the master versions and distribute them to various packages of documents for different users and their uses. As a result, this set of documents is dynamic in a way that is likely to pose problems for long-term maintenance. It is not clear whether archives might be prepared to take responsibility for maintaining a dynamic document set like this.

4.4 RELIABILITY AND ATTRIBUTION. The reliability of a transcription is an important issue, whether it is a transcription of audio or video recordings as commonly made in field research or, as in the Laves case, a retranscription of notes that have been written by someone other than the retranscriber. Reliability is especially important in the Laves case for two reasons. First, as with other historical manuscripts, there are many difficulties in interpreting the notes. This is due mostly to their variable legibility, but it is compounded by the sometimes less than explicit expression found in notes as opposed to more developed writing. Second, care has to be taken so that spurious interpretations are not institutionalized, since there is relatively little documentation of these dialects or of the broader patterns of dialect diversity within Noongar. This section first examines general issues of reliability and then describes the methods used in this case.

We might define the reliability of language description or documentation as an assessment by some individual of the degree to which a user can rely on some materials as an accurate record of language used, in speech or in writing, in the contexts of some place and time in a speech community, and the function served by that language in that community. Under such a broad definition, all written records of spoken language are almost certainly unreliable to at least some degree. The question is how to characterize reliability and even to systematize it, even if only broad degrees can practically be distinguished.

At first glance, the cost of assessing and expressing the reliability of every item in a transcription would not seem justified by the likely benefits. It depends, of course, on the goals of the transcription and the consistency of the material, which together influence the methods to be used. If the current trend in language documentation is towards greater accessibility, achieved by rendering information in a form that can be accessed for different purposes at some later time or by other users, then some form of explicit characterization of reliability should be favored.

If the transcription scheme is designed to allow for automatic processing of the document (for presentation, for producing different versions, or for other information retrieval), this introduces the additional issue of how to maintain the validity of the information by

preventing an item of content from being separated from the reliability information that applies to it. Processing methods will need to take this into account, but this will probably happen only if reliability is adequately documented.

Several questions then arise regarding (1) what kinds and degrees of reliability to distinguish, (2) how to relate reliability assessments to the relevant points in the provenance of an item, (3) whose assessment of reliability it is, (4) where and how to mark it explicitly, and (5) how to retain the reliability assessment through whatever processing the data may undergo.

By the provenance of a transcription I mean the various contributions to the final form, each of which involves its own reliability factors. Here these include the reliability of the information provided to Laves, the reliability of his interpretation of that information and of how he transcribed it, the reliability of the resulting transcription in relation to any physical deterioration, the reliability of the retranscribers' interpretation of these things, and the reliability of their retranscription.

Laves had the benefit of professional training in linguistics so he would probably have been less prone to errors of interpretation than more naïve recorders of that earlier era, and more likely to use a consistent transcription scheme. The types of recording errors that are possible are well documented in the literature on field research (for example Crowley 2007; Vaux and Cooper 2003) and the literature on the interpretation of historical documents (for example Austin and Crowley 1995), and will be familiar to most researchers. These include errors in hearing or analyzing a form, misunderstanding the meaning, simple spelling errors, irregular handwriting, etc.

Laves did not use audio recording in his Noongar work, and since modern recording technology is now a standard tool of fieldwork, especially in the recording of texts, most modern researchers can probably only imagine the limitations incurred. He was presumably attempting to transcribe the stories as they were told to him, and since people generally speak much faster than a researcher can write, there is a clear potential for errors in transcribing. It is likely that speakers had to repeat individual words, phrases, clauses, or larger sections, and in that situation, speakers may tend to give different versions rather than verbatim repetitions. This is presumably because a request for repetition more commonly implies a failure to understand the original version, or because they want to offer something that is easier to write down. We could expect as a result that the transcript might contain omissions or even garbling of different versions. The traditional response to a reliability factor like this is mostly just an implied caveat lector, but there are various ways that this type of reliability factor could be explicitly annotated—for example as <reliability factor= "Presumed direct transcription in interview."/> (as opposed to "transcription from audio-recorded speech" etc.) and indicated in the metadata and/or within the document itself. An appropriate annotation has not yet been applied to the Laves materials.

A separate question also arises where the speaker must repeat sections for the interviewer, because the utterances produced by the speaker may tend to be less natural. How an utterance relates to a speaker's language knowledge or typical use, or that of a speech community, is of course a complex question. Naturalness is not so much an issue of reliability but of being able to characterize utterances in terms of variation in style of speech so that users can interpret these appropriately. Laves gave almost no assessments of this type, apart from characterizing one text as "fragmentary" and noting that one speaker is "very

shaky" on a particular word.18 Interestingly there is also no record of the speaker's own assessments of reliability (for example, as "it might be…" or "I think it's…") or Laves's impression of the degree of reliability that the speaker appears to attach to something. Most of Laves's indications of reliability in the original transcriptions appear to relate to whether he had clearly heard or properly understood something, or sometimes perhaps to the nature of some apparent variation.

Issues of reliability are common at the level of individual items (words, phrases, etc.) in the Laves materials. Degrees of illegibility, whether due to handwriting or to the physical condition of the page, are indicated using <unclear>. Thus, <unclear>qarl</unclear> is to be read that "qarl" most likely represents Laves's intention, but that the named transcriber finds it to be unclear and therefore less reliable. This tag is typically not applied below the level of the word. If the transcriber cannot proffer a reasonably plausible account of a word, it is annotated as <unclear>[unclear.word]</unclear>.

In many cases, however, the transcriber can offer more than one plausible transcription of a word. In addition to which, different transcribers may interpret something differently, or at least have a different set of plausible readings. In example (12) below, two transcribers read the form differently (even though both had developed a familiarity with Laves's handwriting style). The second character could plausibly be read as either "u" or "o." This is annotated, as in (13), as alternative transcriptions within a set, the <alt_set>. Each of the competing transcriptions is attributed to a specific individual: the transcriber of the first version here is understood as the transcriber identified for the page as a whole. This follows an important principle that the immediate source of information should always be identified together with the time of its incorporation into the body of data.

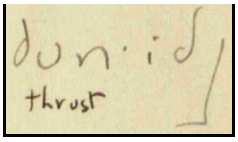

FIGURE 12: Example of ambiguity in Laves's hand-writing.

(13)

```
<alt_set>
<alt>don:idj</alt>
<alt transcriber= "HW" date transcribed="010303">dun:idj</alt>
</alt_set>
```

This structure is intended primarily to provide a way to ensure that all the plausible interpretations are represented and does not attempt to formalize any relative or absolute specification of the reliability of any of them. In the transcriber's practice, there is a weak

[18] Laves indicates when a text is of lesser quality by appending one or more Qs to the otherwise numerical identifier for each text, for example "Text 170QQQ," but he does not specifically relate this to naturalness.

implication that the most plausible is given first. In the default facsimile version, the first item appears in the text on the facsimile page and the alternatives are placed in transcriber's footnotes. The <alt_set> structure is used mostly for single words but can be used to represent alternatives at any level of structure, including, for example, different interpretations of the alignment of words and glosses in interlinear glossing. The TEI P5 Guidelines have a number of possible equivalent functions—*choice*, *alternation group*, and even *apparatus* elements—and a more articulated scheme of specifying the likelihood of a given alternative which even permits a numerical expression of probability (Burnard and Bauman 2007).

The other method for representing alternative interpretations of form or meaning is the traditional one: transcriber's annotations to the text. The implementation here explicitly indicates the domain of the annotation by marking both the beginning and the end of the section of the text to which the note refers, rather than the common practice of placing the footnote marker at the end of the relevant section. The content of the transcriber's notes is not formally limited, but in practice they relate mostly to forms, meaning, structure of the text, or even expansions of abbreviations, as the following examples illustrate.

(14)

```
<transcriber_note_domain>Kurinj</transcriber_note_domain>
<transcriber_note>Could be 'rr' originally and overwritten
as 'r'.
       </transcriber_note>
```

(15)

```
<transcriber_note_domain>uncle </transcriber_note_domain>
<transcriber_note>Perhaps this should be mother.</
transcriber_ note>
```

(16)

```
<transcriber_note_domain>… </transcriber_note_domain>
<transcriber_note>Uncertain as to whether this belongs to Text 174
or
          175.</transcriber_note>
```

(17)

```
<transcriber_note_domain>bro-law </transcriber_note_domain>
<transcriber_note>brothers-in-law</transcriber_note>
```

Laves made heavy use of special symbols and abbreviations for linguistic and other information. He provided a gloss for seventy-odd symbols, though the descriptions are not always particularly explicit. Example tokens are given in figure 18. Each is indicated in the retranscription as a custom XML Entity, a variable specified in the XML retranscription for the purposes of this project for which some text or a graphic image can readily be substituted for presentation.

	Laves's description	**XML Entity**
	onomatopoeic or possibly so	&onomat;
	abbreviated form [of] word or sentence	&abrvd;
	prefix	&prfx;
	contextual (not contained in passage)	&ctxl;
	questionable record/meaning	&qstnbl;

FIGURE 18: Examples of Laves's documented special symbols.

There are two reliability issues in relation to the special symbols. First, there are many other symbols for which Laves does not provide a meaning. The consequences of having lost this information depend on what type of information is involved. For some types, like the "onomatopoeia" symbol above, the meaning of the symbol could be lost while the text content remains as an independent item. However, if a symbol modifies the meaning that is expressed, like the "questionable" symbol above, the loss of that information would lead to a misinterpretation of the reliability that Laves attached to the text. If readers do not know the meaning of a symbol, they are likely to rely on the text and to ignore the possible contribution of the symbol, and therefore to possibly overestimate the significance of the written word. One such undocumented symbol, which is identified as "&unknown3;" in the XML re-transcription, and illustrated in figure 20 below, is very common in the notes and therefore constitutes a significant problem of interpretation. Attempting to infer its meaning across the range of contexts that it appears in, suggests meanings like: "query," "possibly," "presumably," "check this," or "alternative(ly)."

The second reliability issue arises because the symbols are handwritten and vary in form from token to token, as illustrated for the documented and undocumented symbols in

figures 19 and 20. The problem is that it is sometimes difficult to determine whether a given instance is an allograph of a known symbol grapheme or a distinct grapheme. The problem is compounded because Laves used some similar looking symbols for distinct categories. Compare, for example the "onomatopoeic" and "abbreviated" symbols above. As a working principle for the retranscription, if a symbol was not clearly an allograph of a known grapheme, it was treated as a distinct grapheme. The result is that the long list of symbols[19] used in the retranscription includes quite a number with only one or two instances. In some cases, the retranscribers' documentation indicates a possible relationship to another symbol grapheme, but no formal mechanism was implemented in the XML to represent this.

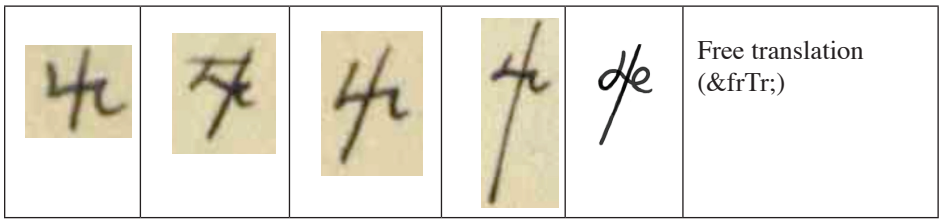

FIGURE 19: Variation in form of documented special symbol

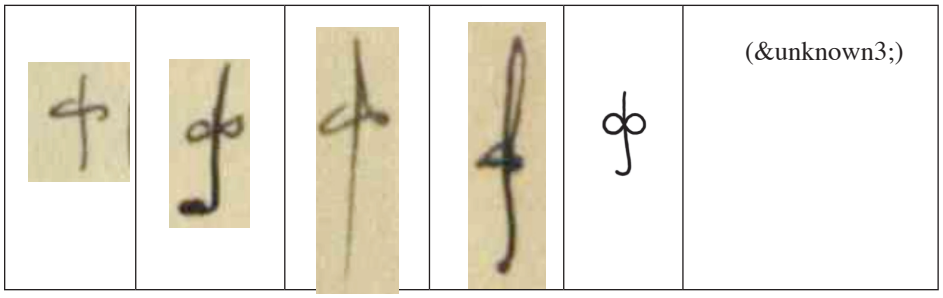

FIGURE 20: Variation in form of undocumented special symbol.

It is not clear that these problems of interpretation are best solved as part of the initial retranscription as opposed to subsequent analysis, and for this reason it was decided that these symbols would simply be identified without any specific interpretation beyond what Laves provided. In order to maintain the facsimile representation in the facsimile versions, the more common symbol graphemes are graphically reproduced. But rather than reproducing the specific allograph used at a given point, a redrawn standard form is used, as illustrated in figures 19 and 20 above. The redrawn forms are intended to capture

[19] A careful review at this stage might well reduce the number of distinct symbols that should be recognized.

the basic form of each symbol grapheme and to be visually distinct from each other. The less common symbol graphemes are currently represented in the facsimile versions by an abbreviation based on the XML entity assigned to them, for example "Symb[grmTns]" for "&grmTns;". The current presentation of these symbols could be improved by indicating the meaning of a symbol, where known, at each instance of that symbol in the facsimile version, perhaps in the status bar or in a pop-up text in the browser window, in addition to documenting them in the accompanying materials.

Custom XML entities were also initially used within the actual retranscriptions to represent the fairly limited set of special phonetic symbols that Laves used in his fieldnotes, for example [ɣ] is represented as "&g1;" and schwa as "&e1;". These entities are then equated with their Unicode equivalents within the document type declaration at the start of the XML document by means of declarations of the form <!ENTITY g1 "ɣ">. This allowed the documents to be used on different platforms and with different software which did not (consistently) implement Unicode. It also facilitates entering data because the entity labels are descriptive to a degree. If a change of symbol was necessary, a change could be made to a single entity declaration which then applied automatically to all instances in the document. All entities are documented separately with whatever is known about the symbols as used by Laves.

4.5 GRAPHIC FORM AS CONTENT. For the reasons given above, the XML scheme used in this project is designed to allow for facsimile representation of the pages, and therefore needs to be able to represent various aspects of graphic form, including the position of text elements on the page. An important question is the degree of accuracy that is needed in order to produce the benefits of facsimile presentation. If it is done in a very high degree of detail, it would be possible to recreate from the retranscription a high-fidelity facsimile version where a human reader would make the same interpretations from the page layout as those available in the originals, in which case the option of access to the original images would offer the reader no advantage.20 However this approach can be relatively costly in terms of the complexity of the XML markup scheme, and therefore the cost of annotation, as well as the complexity of the XSLT stylesheets needed and the costs of developing them. An alternative strategy is used in this project—that is, to encode a lesser degree of detail that captures only the more important visual aspects of the page, and those only to an approximate degree of precision. For a user interested in the interpretation of some specific content on a page, the rough facsimile can readily be augmented by viewing the image of the original page.

---

[20] The gold standard would be a scheme that precisely mapped each retranscribed word or graphic element to its source on the photographic image of the page. An example of a scheme of this type can be found in Google Book Search (http://books.google.com/), where search results visually highlight the user's search terms directly on the scanned image of the original page. This is presumably feasible only because they are using typeset source materials and the automatic optical character recognition that high-quality lettering permits.

Specific tags are used for the more structured aspects of graphic form, such as headings and interlinear glosses.[21] Less structured or less conventionalized aspects are described in transcriber's annotations. The page shown in (21) shows features of both types. In addition to the obvious interlinear glossing, there is double underscoring both beneath a text word and apparently as a marker in the margin, circling of a gloss, and a (faint) arrow connecting two items on different lines. (All in a contrasting color to the text.)

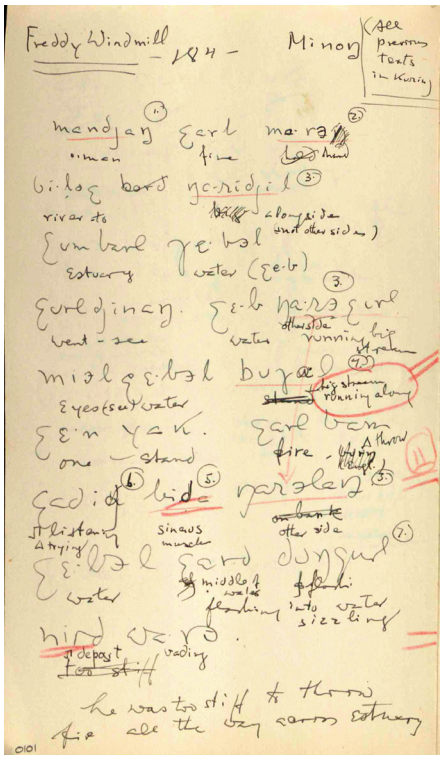

FIGURE 21: Notebook page illustrating various aspects of graphic form.[22]

Absolute position on a page is schematized in a page model (or template) on the basis of the more consistent patterns found in the original field notes. Two page models are sufficient for the Laves materials—one for the notebook pages, which are in portrait orientation, and one for the loose slips, which are in landscape orientation. The page model for slips is shown in (22).

---

[21] It must, of course, be recognized that interlinear glosses are related not only by a conventionalized relative position but also by the semantic relation between an item and its gloss.

[22] This page is reproduced here by kind permission of Ezzard Flowers on behalf of the family of Freddy Winmer, the author of the text from which it is taken.

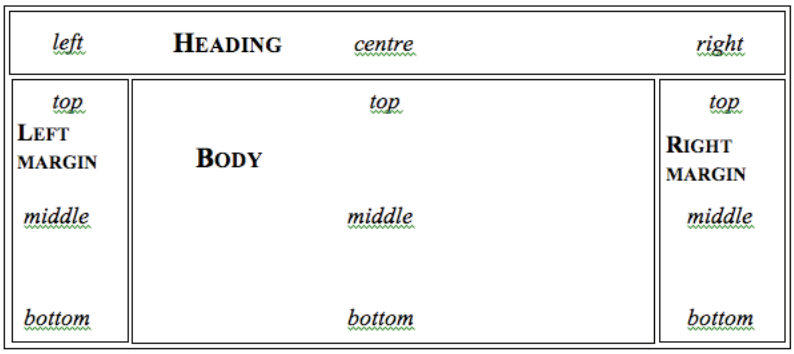

FIGURE 22: Page model for loose slips.

In the XML scheme, the main components of the page model (HEADING, BODY, LEFT MARGIN, and RIGHT MARGIN) are represented as distinct tags and the positions within each component as values of the *pos(ition)* attribute, as illustrated by the edited example in (23). In practice, the center value is the default for a heading, and the top value is the default for the other components.

(23) Example of page layout.

```
<image>
       <heading pos= "centre">Types</heading>
       <heading pos="right" mode="stamped">KURIN</heading>
       <leftmargin pos= "top">cf. 207-21</leftmargin>
       <leftmargin pos="bottom">6</leftmargin>
       <body >…</body>
</image>
```

Relative position on a page is relevant in interlinear glossing, column and table formatting, and in diagrams. The markup of interlinear glossed text has been discussed above. It is rare in the materials for the entire body of a page to be divided into columns, but it is relatively common in the slips for fairly compact notes to appear in a format that can be represented in columns. This graphic form is represented using the column set (<col_set>) and numbered individual columns within it. Content that is clearly in tabular form is represented using HTML-like table structure (<table>,<tr>,<td>, etc.) in the XML, but this is also used for content where the relative positioning is merely table-like and can be efficiently represented in this way.[23] In practice, table-like and column-like relative posi-

[23] This parallels the traditional use of XHTML table structures to represent the arrangement of layout blocks in web page design.

tioning on the original page are not always easily distinguished, and because of this there is some inconsistency in the retranscription. However this is not a crucial issue, because column form is equivalent in graphic form to a table with a single row, and the markups involved are readily interchangeable if necessary.

4.5.1 DIAGRAMS. Diagrams are potentially the most difficult case of graphic form as content. The Laves materials contain twenty-odd genealogical diagrams and just a few others. The genealogical diagrams contain the usual marriage and descent information, together with an attempt to record traditional moiety membership and totemic affiliations. Many extend over two pages of the notebook volumes. For the retranscription, the diagrams were analyzed and represented using the GEDCOM XML scheme (LDS 2002), which is based on genealogical information types rather than the actual graphic elements of the diagrams. Because it was difficult to transform this XML into a diagram for presentation purposes, the diagrams also were redrawn in a legible form using diagramming software and stored in PDF format.24 The non-genealogical diagrams are diverse and do not follow any conventionalized structure. See (24) for example. No attempt is made to represent their graphic form directly in XML. A brief description is given in a special transcriber's annotation, using <diagram> tags, but at this stage users will generally need to consult the image of the original diagram.

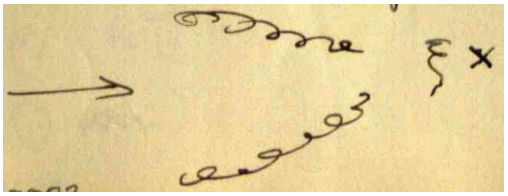

FIGURE 24: Nonce diagram accompanying text.

5. CONCLUSION. In this paper I have described the goals and parameters of a project to digitize Gerhardt Laves's handwritten Noongar field notes, and have discussed key areas of the implementation of a retranscription in XML. The paper demonstrates that the language documentation process follows from the nature of the material, together with the goals set for the usability of the content, as negotiated with the language community. The goal of providing a resource for community language activities has partly been met: edited versions of some of the texts have since been prepared for publication in a community project. Other similar work has been mooted. A short select word list drawn from the notes has been prepared for the Noongar families for their use and to familiarize them with the potential value of the notes. Their value for linguistic analysis has been initially demonstrated by a sketch grammar prepared as a student project (Wykman 2005), and it is hoped that the notes will be further exploited for more extensive analysis in the future.

[24] For the purposes of community discussions on rights in the Laves materials, a hardcopy booklet was produced showing, for each genealogical diagram, the original image and the redrawn diagram at an opening.

Appendix

Excerpt of XML retranscription of page in (1) and (3).

```
   <imageset>
   <version date_of_extraction_from_volumesMaster="Monday, 6
November 2006 9:48:58 AM"/>

   <vol number= "23">
   …
   <text id= "166" author="Freddy Winmer or Windmill">
   …
   <text_proper id="166" author="Freddy Winmer or Windmill">
   …
   <image
   …
   filename="m0037456_i2.23_v_p0093_m"
   volume= "23"
   page= "0093"
   transcriber="HW"
   date_transcribed="091203">

   <heading  pos="centre"></heading>
   <heading  pos="left"></heading>
   <heading  pos= "right"></heading>

   <body >
     <gl_set>
       <gl_pair>
         <glossed_line >
           <glossed>nu&ng;arl</glossed>
           <glossed>ni&dotn;t</glossed>
           <glossed>barda&ng;</glossed>
           <glossed>bara&ng;</glossed>.
         </glossed_line>
         <gloss_line>
           <gloss>man</gloss>
           <gloss>tail</gloss>
           <gloss>&lit; jump - </gloss>
           <gloss>seize</gloss>
         </gloss_line>
       </gl_pair>
       <gl_pair>
         <glossed_line >
           <glossed>qad</glossed>
           <glossed>bindilj bindilj</glossed>
           <glossed>bam</glossed>.
         </glossed_line>
         <gloss_line>
```

```
            <gloss>head</gloss>
            <gloss>battering, smashing</gloss>
            <gloss>hit</gloss>
            </gloss_line>
        </gl_pair>
        <gl_pair>
          <glossed_line >
            <glossed>nirnt</glossed>
            <glossed>burn</glossed>
            <glossed>bara&ng;</glossed>.
            <glossed>qa:daq</glossed>
          </glossed_line>
          <gloss_line>
            <gloss>tail</gloss>
            <gloss>cut</gloss>
            <gloss>seize</gloss>
            <gloss>head</gloss>.
          </gloss_line>
        </gl_pair>
        <gl_pair>
          <glossed_line >
            <glossed>nirnidj</glossed>
            <glossed>day&E;&rr;&e1;<annot           place="super"
type="footnote ref">&circ1.;</annot></glossed>.
            <glossed>qo:&rr;</glossed>
            <glossed><unclear>ni</unclear></glossed>
          </glossed_line>
          <gloss_line>
            <gloss>deposit</gloss>
            <gloss>
              <col_set>
                <col1>dog's tail<br/>
                  <annot  type="strikeout">&ctxl; &frTr; like a
king's crown</annot>
                </col1>
              </col_set>
            </gloss>
            <gloss>back</gloss>
            <gloss></gloss>
          </gloss_line>
        </gl_pair>
      </gl_set>
      <gl_set>
        <gl_pair>
          <glossed_line >
            <glossed>qumba&rr;&e1;</glossed>
            <glossed>mai</glossed>
            <glossed>wa&ng;qa&rr;&e1;</glossed>
          </glossed_line>
          <gloss_line>
```

```
            <gloss>much</gloss>
            <gloss>sound</gloss>
            <gloss></gloss>
          </gloss_line>
        </gl_pair>
        <transcriber_note_domain>
        <g_translation>&frTr; The fellow talks a lot : important</g_
translation>
        </transcriber_note domain>
        <transcriber_note>Horizontal lines extend outwards from free
translation,
        apparently to indicate its application to the whole preceding
line.
        </transcriber_note>
     </gl_set>
     <gl_set>
        <gl_pair>
          <glossed_line >
            <glossed>nu&ng;a&rr;</glossed>
            <glossed>barda&ng;<annot place=“super” type= “footnote
ref”>&circ2.;</annot>
            </glossed>
            <glossed>dandi    bi&rr;i    b<annot    place=“super”
type=“footnote ref”>
                  &circ3.;</annot>i&rr;a
               <alt_set>
                  <alt>nj</alt>
                  <alt>&ng;</alt>
               </alt_set>
            </glossed>
          </glossed_line>
          <gloss_line>
            <gloss>man</gloss>
            <gloss>
                 <col_set>
                    <col1>
                       <annot      type=”strikeout”>big      mob</
annot><br/>
                       walking
                    </col1>
                 </col_set>
            </gloss>
            <gloss>
                 <col_set>
                    <col1>&frTr; gathered together<br/>
                       &frTr; mob alongside him<br/>
                       all under him<br/>
                       people all huddled<br/>
                       together
                    </col1>
```

```
                </col_set>
            </gloss>
          </gloss_line>
        </gl_pair>
      </gl_set><br/>

   - Finis of <annot place="super" type="insert">166</annot><annot
type="strikeout">230</annot> - <br/>
   <section_break type="horizontal"/>

   </body>
   <body pos="bottom">0093</body>
   </image>
   …
   </text_proper>
   …
   </text>
   …
   </imageset>
```

References

Austin, Peter, and Terry Crowley. 1995. Interpreting old spelling. In Paper and talk: A manual for reconstituting materials in Australian indigenous languages from historical sources , ed. by Nicholas Thieberger, 53–102. Canberra: Aboriginal Studies Press.

Bindon, Peter, and Ross Chadwick. 1992. A Nyoongar wordlist from the southwest of Western Australia. Perth: Western Australian Museum.

Bird, Steven, and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. Language 79(3): 557–582.

Bow, Cathy, Baden Hughes, and Steven Bird. 2003. Towards a general model of interlinear text. Paper presented at EMELD 2003, Michigan State University.

Brandenstein, C.G. von. 1988. Nyungar anew : Phonology, text samples and etymological and historical 1500-word vocabulary of an artificially recreated Aboriginal language in the south-west of Australia, Series C. Canberra: Pacific Linguistics.

Burnard, Lou, and Syd Bauman, eds. 2007. The TEI guidelines. [Guidelines for electronic text encoding and interchange]. P5 ed. Charlottesville, VA: Text Encoding Initiative Consortium. http://www.tei-c.org/Guidelines/P5/.

Burnard, Lou, and C. M. Sperberg-McQueen. 2006. TEI Lite: Encoding for interchange: An introduction to the TEI Revised for TEI P5 release. http://www.tei-c.org/release/doc/tei-p5-exemplars/html/teilite.doc.html.

Crowley, Terry. 2007. Field linguistics: A beginner's guide. Oxford: Oxford University Press.

Dench, Alan. 1994. Nyungar. In Macquarie Aboriginal words, ed. by Nicholas Thieberger and William McGregor, 173–192. Sydney: Macquarie Library.

———. 2000. Comparative reconstitution. In Historical linguistics 1995. Volume 1: General issues and non-Germanic languages, ed. by John C. Smith and Delia Bentley, 57–72. Amsterdam: John Benjamins.

Douglas, Wilfrid H. 1968. The Aboriginal languages of the south-west of Australia. Canberra: Australian Institute of Aboriginal Studies.

Greetham, David C. 1994. Textual scholarship : An introduction. New York: Garland.

Haugen, Odd Einar, ed. 2008. The Menota handbook: Guidelines for the electronic encoding of Medieval Nordic primary sources, Version 2.0. Bergen: The Medieval Nordic Text Archive.

Henderson, John. 2006. Right(s), permission(s) and protocol(s) in language documentation: Laves' 1931 Noongar field notes. Presented at Hans Rausing Endangered Languages Project Seminar, School of Oriental and African Studies, University of London.

Henderson, John, Andrew Gargett, David Nash, and Denham Harry. 2003. Interpretation and re-presentation of historical language materials: Laves' 1931 Nyungar notes. Paper presented at the 2003 Conference of the Australian Linguistic Society, University of Newcastle.

LDS, Church of Jesus Christ of Latter-day Saints Family and History Department. 2002. GEDCOM XML Specification Release 6.0. Salt Lake City: Church of Jesus Christ of Latter-day Saints.

Nash, David. 1993. Gerhardt Laves 15/7/1906–14/3/1993 [Obituary]. Australian Aboriginal Studies 1:101–2.

O'Grady, Geoffrey N., Carl F. Voegelin, and Florence M. Voegelin. 1966. Languages of the world: Indo-Pacific fascicle 6. *Anthropological Linguistics* 8(2): 1–197.

Scott, Kim, Hannah McGlade, Denise Smith-Ali, and John Henderson. 2006. A protocol for Laves' 1931 Noongar field notes. Perth: University of Western Australia.

Vaux, Bert, and Justin Cooper. 2003. *Introduction to linguistic field methods*. Munich: Lincom Europa.

Wykman, Harry. 2005. A description of Kurin/Minong Noongar as documented by Gerhardt Laves in his field notes of 1931. University of Western Australia PhD dissertation.

John Henderson
john.henderson@uwa.edu.au